THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Margaret Evangeline Froggatt

TITLE OF THESIS STUDENT EVALUATION IN A JUNIOR HIGH SCHOOL: A COMPARISON OF TEACHER-MADE AND STANDARDIZED MEASURES.

DEGREE FOR WHICH THESIS WAS PRESENTED MASTER OF EDUCATION

YEAR THIS DEGREE GRANTED Fall 1984

DATED: October 11, 1984

THE UNIVERSITY OF ALBERTA

STUDENT EVALUATION IN A JUNIOR HIGH SCHOOL: A COMPARISON OF TEACHER-MADE AND STANDARDIZED MEASURES.

by

Margaret Evangeline Froggatt

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF EDUCATION

IN

Counselling

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

FALL 1984

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled STUDENT EVALUATION IN A JUNIOR HIGH SCHOOL: A COMPARISON OF TEACHER-MADE AND STANDARDIZED MEASURES submitted by Margaret Evangeline Froggatt in partial fulfilment of the requirements for the degree of MASTER OF EDUCATION in Counselling.

Date 9 October, 1984

ABSTRACT

The primary purpose of this research was to determine the degree of relationship between teacher-made and standardized measures of achievement.

Report card scores, which reflected teacher-made test results were collected in language arts, mathematics, science and social studies. These scores included November, 1981, March, 1982 and June, 1982 results for Grade 7, 8, and 9 students. Teacher rank orderings of students of expected year end performance were also compiled. These two measures were compared both to each other and to the Canadian Achievement Tests for all three grades. The Differential Aptitude Test results were also compared for Grade 9 students.

This correlational study found surprisingly high correlations between the report card marks. Expected correlations were found between the standardized tests and ranged between .511 and .821 on content similar subtests.

The results of the study led the author to conclude factors other than curriculum content were influencing teacher grading. These may have been teacher attitude or expectations which, in fact, were not measured in this study.

ACKNOWLEDGEMENTS

A special thank you is expressed to Dr. George Fitzsimmons for his encouragement, support and faith in me.

Thank you to Dr. Peter Calder and Dr. Ken Ward for serving as members of my committee.

Appreciation is extended to the 1981-82 Jubilee Junior High School staff and principal, Dr. A.O. Jorgensen. Without their contributions, the research could not have been carried out.

My gratitude is expressed to my "family of friends" who cheered me on during the last year especially Marg, Donna, Norma, Dave and Lydia. Dr. Sylvia Gorchynski's excellent support was also much appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. Introduction

Classroom teachers get to know their students through numerous and varied means. Direct daily observation provides a wealth of information. Interactions with the students in academic activities adds more to the growing body of knowledge. Informal teacher-student participation in games or social activities may supply a different perspective on the child. Evaluation by means of teacher-made tests continues to add data to the teacher's knowledge of the student.

As this process of data collection continues, teachers often become aware of their personal biases and expectations that are becoming attached to the knowledge about the child. They begin to ponder the accuracy of their assessments. Questions related to the validity of the classroom tests begin to arise. The staff wonder if they are able to trust these exams to be measuring the content of the curriculum.

Standardized test results can also add substantial information about the student. However, after these tests are purchased, administered, scored and student profiles are returned to the school, the results may sit unused. Added to this possibility, is the general overall question of whether standardized tests provide enough useful information to make them worthwhile administering.

The purpose of this research came about in an attempt to explore a number of questions surrounding the issue of teacher-made and standardized test measures. As a guidance counsellor, the writer recognized the importance of standardized measures being able to assist

both teacher and student in knowledge related to pupil progress and achievement. Later, when the role of the writer changed to that of an assistant administrator, standardized measures became important also in looking at program evaluation and comparing the results of our school population to those of other students across the country.

Nunally's (1975) comment concerning tests being useful only if they assist in making decisions related to the student's progress and ability, the curriculum and the method of instruction used in the classroom, became the underlying issue for this research. Were standardized tests supplying useful data in these three areas? Could teacher-made evaluation measures provide just as much information as standardized tests did?

These questions were explored through collection of various data at Jubilee Junior High School in Edson. Teacher rank-orderings of students, teacher-made test results as reflected in three report card grades, results of the Canadian Achievement Tests (CAT) and the Differential Aptitude Tests (DAT) were collected and correlated.

Four questions have been addressed.

1. What kind of relationship exists between grades based upon teacher-made tests and the independence of those grades among four core subjects?
2. Does a link exist between teacher rank-ordering of students and 1) their final grades, 2) the Canadian Achievement Tests, and the Differential Aptitude Tests?

3. How will student year end grades correlate with the CAT and the DAT?

4. What is the relationship between the CAT and the DAT?

Consideration of the relationship of these measures may provide useful information which can supplement the decision making processes surrounding student evaluation.

II. REVIEW OF THE LITERATURE

A. Introduction

Schools are decision making centers. The onus falls upon administrators, teachers and counsellors alike to make sensible, well-informed decisions related to the many educational issues which affect their students. Student evaluation is one of those issues which is surrounded by numerous and complex decisions and certainly draw on the creative energies of a school staff. How all the related aspects of evaluation are dealt with in a school can greatly effect the climate and tone of that establishment.

In this chapter, the writer reviews seven components of the decision making process as it relates to student evaluation. Initially, the global concept of school testing programs is explored followed by some general standardized test information. This is enlarged on in a more specific look at achievement tests. Teacher-made tests, rank-ordering of students within classrooms and the judgement laden area of assigning grades are then considered. The chapter concludes with a brief overview of the effects teacher expectations have on achievement.

B. School Testing Programs

School testing programs have long supplemented in-school teaching and learning. Such programs involve the administration and utilization of the results of various types of standardized tests such as achievement, aptitude and intelligence tests and interest inventories.

For any testing program to be effective, it must serve the needs of the specific school it is used in. As a result, the program should be

integrated with the goals and educational objectives of that school (Ahmann & Glock, 1981).

Shertzer and Linden (1979) found that testing programs which provide "measurement data, are indispensible in evaluating the progress of students toward educational goals, and they can aid in making administrative decisions concerning the appropriate classification of students" (p. 481).

They further advocate that testing programs must be administered at a time in the school year when maximum use can be made of the results. Fall testing dates have become considerably more popular as schools recognize the need for "outside" information on the students progress. Teachers, administrators and counsellors must be made aware of the results and have previously made a joint commitment to use them toward furthering the schools goals. This commitment ensures that the testing program be integrated into the entire education program of the school. A testing program cannot be viewed as an end in itself. Regular use of the selected tests will ensure the program continuity both in progress and growth. However, it is also essential that systematic evaluation and change, as is necessary, be part of the system.

Ebel (1979) suggests that a school testing program would supplement teacher-made classroom achievement tests with school-wide testing in order to

1. provide information needed for instruction and guidance
2. evaluate local school achievement against external standards

3. stimulate and direct continuing efforts to improve curriculum and instruction in the local school (p. 319).

Jubilee Junior High School in Edson could provide an illustration of a school testing program which incorporates teacher-made tests with several standardized tests.

Historically, the Canadian Tests of Basic Skills (CTBS) had been administered to students at Jubilee. However, as the Alberta curriculum changed, the CTBS content no longer matched the provincial requirements. The teaching staff seldom used the test results. Consequently, use of the CTBS was discontinued about 1979.

In 1980-81, the school participated in a norming study conducted on the Canadian Achievement Tests (CAT). During the 1981-82 school term, the CAT became the formal achievement battery in Jubilee's testing program. Standardized achievement tests were the only tests administered to the grade 7 and 8 students. As well as the CAT, the grade 9 classes completed the Differential Aptitude Tests (DAT) in January of 1982 and the Kuder Interest Inventory in March of the same year.

Teacher-made tests were completed within individual classes according to course outlines and as required. However, major teacher-made tests corresponded with set report card dates.

The following timetable outlines the testing program at Jubilee Junior High in 1982.

STANDARDIZED TESTS		TEACHER-MADE TESTS
	September (1981)	Teacher-made tests
	October	as required *
	November	Report Card Specific Exams
	December	*
D.A.T. (Grade 9)	January (1982)	
	February	
Kuder Interest Inventory (Grade 9)	March	Report Card Specific Exams
	April	*
CAT (Gr. 7,8,&9)	May	
	June	FINAL EXAMS

This standardized testing program served the needs of the school to a considerable degree. It provided good information, especially at the grade 9 level in terms of the student's aptitude and interests. This, combined with the student's achievement as reported by teacher-made tests and grades provided students, teachers and parents with invaluable information related to helping make high school and career choices.

It would be interesting to compare the principles outlined by Shertzer and Linden, as well as those of Ebel to the standardized testing program in effect at Jubilee.

Various types of standardized tests used in school testing programs will now be surveyed.

C. Standardized Tests

Hills (1981) defines a standardized test as one "which is always given under the same conditions so that scores can be compared across groups" (p. 136). Other criteria which Hills suggests constitute a standardized test include: test scores being reflected in norms, administration of the test to large groups under timed conditions, objective scoring, test construction completed by professionals, and publishers providing access to all test material including manuals and test reports.

Nunnally (1975) tells us "norm literally means average, and statistical norms consist of comparing scores of individuals in a group with the average response in the group" (p. 10). Norms are obtained by accumulating test scores from a specified sample and converting them to derived scores such as percentile ranks, grade equivalents, stanines and so on. It is important to realize that "a comparison of scores with norms is meaningful only in the context upon which the norms are based" (p. 124). In other words, the norm group must be similar to the tested sample.

Standardized tests can provide school staff with information not readily available from teacher-constructed tests or classroom interactions. Aptitude, intelligence, scholastic achievement, vocational interests and personality characteristics are some areas standardized tests are proficient in measuring.

This study will focus on the use of two types of standardized tests: achievement and aptitude tests. There are several important distinctions which must be recognized between the two.

Anatasi (1982) defines achievement tests as measuring the effects of relatively standardized sets of experiences such as knowledge of basic skills related to specific curriculum content. She further states "aptitude test performance reflects the cumulative influence of a multiplicity of experiences in daily living" and "serves to predict subsequent performance" (p. 393). In estimating the validity of these two types of tests, a very basic difference becomes apparent. Achievement test measures are usually evaluated in terms of their content validity, while predictive criterion-oriented validity is the most characateristic way of assessing aptitude tests.

Four important differences between aptitude and achievement tests are:

1. Scholastic aptitude tests sample more broadly
2. Scholastic aptitude tests sample outside learning as well as school subjects. Achievement tests focus on a particular academic curriculum.
3. Achievement tests sample more recent learning.
4. Scholastic aptitude tests predicts future performance, while the achievement tests seek to measure academic progress. (Cleary et al., 1975, p. 21. As cited in Ahmann & Glock, 1981, p. 285).

Through an achievement test we may have an indication of what a child knows at present, while an aptitude test will tell us what his or her future performance may be like. However, warnings abound from numerous sources as to the dangers of applying these definitions too rigidly.

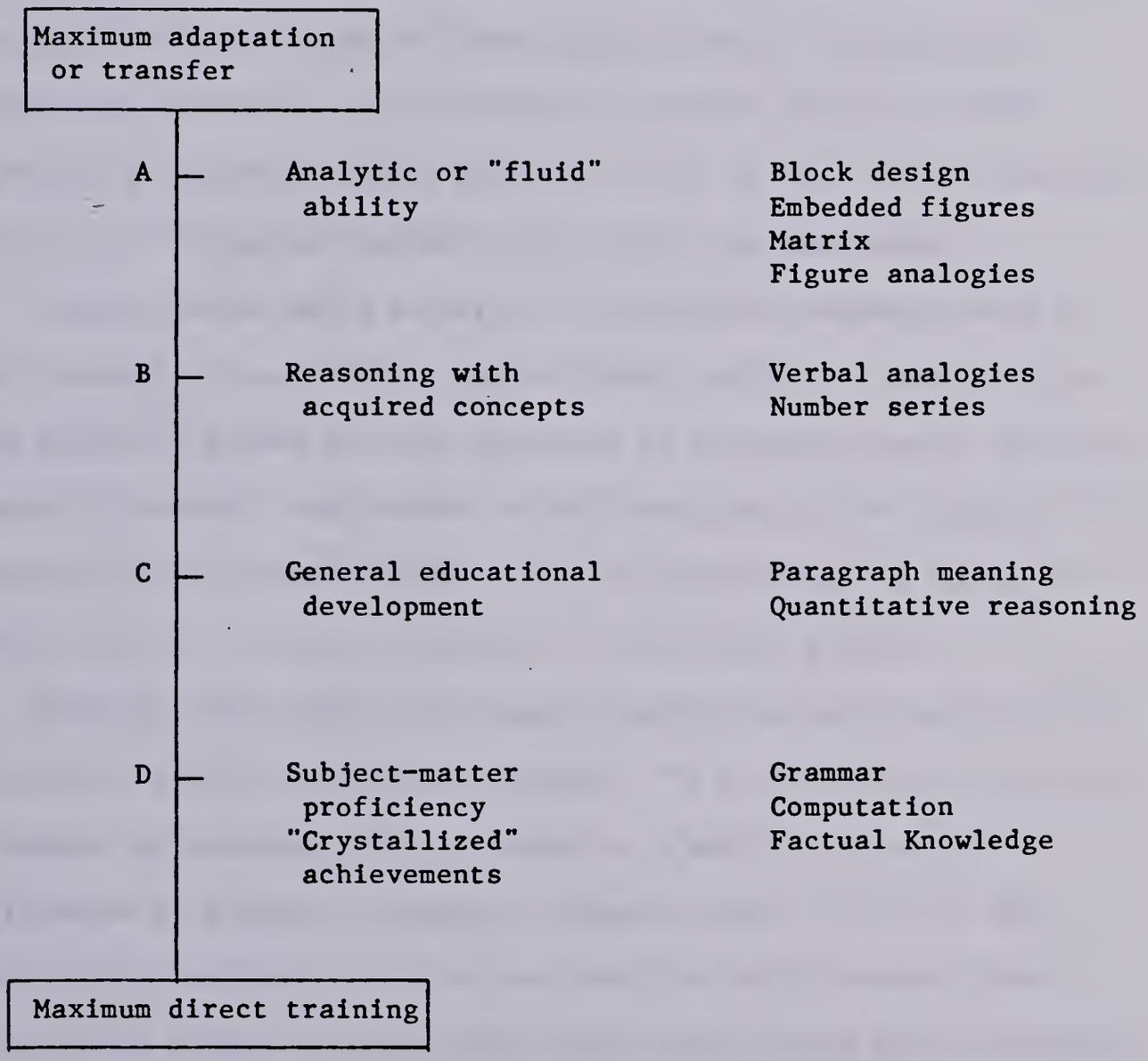
Cronbach (1970) writes "some experts contend that we are only fooling ourselves when we call some tests "achievement" tests and others "aptitude" tests, because the two really measure the same thing" (p. 283). Using the Differential Aptitude Test (DAT) as an example, it is apparent that this comment holds. Cronbach's "spectrum for comparing tests of scholastic aptitude or general ability" (p. 282) would place DAT subtests in language usage and spelling at the "D" end where "maximum direct training" and "subject matter proficiency" influence the students' test scores. [SEE FIGURE 1]

The remainder of the DAT subtests would likely fall within the "C" categories. The CAT would also be included within the "C" and "D" direct training end of the spectrum.

Tests administered within a school that would lie in the "A" and "B" range of Cronbach's spectrum would be group and individual intelligence tests. The group tests may be administered by the classroom teachers, however, the individual measures must be administered and interpreted by a qualified counsellor or school psychologist.

The 5th edition manual of the DAT, relays the correlation coefficients of DAT scores and scores of the test of a number of

FIGURE 1 Spectrum for comparing tests of scholastic aptitude or general ability



Note: From Essentials of psychological testing (p. 282), by L.J. Cronbach, 1970, New York: Harper & Row. Copyright 1970 by Lee J. Cronbach.

achievement batteries. The relationships are seen as "substantial" (p. 138) by Bennett, Seashore and Wesman (1974).

Using the SRA Achievement Series (Form D), subtests in arithmetic, language arts and reading for both girls and boys in grade 9, the correlation coefficients with the DAT for the VR, NA and VR + NA subtests, given concurrently, ranged from .51 to .81. Correlations for three other achievement measures fell within the same range.

Another factor which aptitude and achievement measures share is their dependency upon reading and arithmetic skills. A student with weak skills in either of these areas may be penalized when an aptitude measure is sought. Test scores in both areas may also be affected by a students attitude toward school where the measures may be viewed as "another test" - to either stimulate or discourage progress.

Generally then, both tests measure aspects of achievement and can be used as predictors of future success. "If a student does well on an arithmetic achievement test, the score is a good predictor of performance in algebra or geometry" (Ahmann & Glock, 1981, p. 284). Furthermore, aptitude tests "are very similar to achievement tests because they measure knowledge and skills that require direct training" (p. 285).

While it is important to recognize the similarities of aptitude and achievement tests, it is essential that the functions of the two not be confused. Achievement tests focus on specific curriculum content related to what the student is in the process of learning. The results can be used in the classroom to evaluate the students

weaknesses and strengths at the end of a training period, then plan a program to meet the needs indicated. Aptitude tests are used to predict future performance. They can be used to help the student to plan for future high school and post secondary courses. Applications of the test results for these two types of tests must be utilized in the appropriate context.

While these are some of the general issues in the area of standardized testing, at this point more specific elements in the area of achievement testing will be examined.

D. Achievement Tests

Gronland (1981) notes that standardized achievement tests fall into two categories: norm-referenced and criterion referenced tests. The former "measure a pupil's level of achievement in various skill and content areas by comparing his/her performance with the performance of other pupils in some general reference group" (p. 303).

Criterion-referenced achievement tests are relatively recent publications and

> are typically interpreted by expressing the degree to which the content or skill representing a particular instructional objective has been mastered. Thus, they describe in terms of specific educational tasks what each pupil has learned and what he/she has yet to learn in some clearly specificed achievement domain (p. 303).

The majority of standardized achievement tests which dominate the educational field are norm-referenced. The criterion-referenced tests

were developed to measure more specific details related to mastery learning and individualized instruction.

There is a trend in test development to attempt to combine both norm and criterion-referenced tests. This will be discussed later in this section.

Achievement test score results, if used effectively can supplement what is already known about the student (Goldman, 1969). With this increased knowledge, decision making can be enhanced to determine if further assessment such as diagnostic tests need to be implemented. This may then result in decisions being made regarding specialized placements, individualized or remedial instruction (Ebel, 1979).

Linn (1983) further suggests that program evaluation is another possible use for these test results.

> The principal who focuses on the results may increase the instructional importance of standarized tests and so may the heavy reliance on the results for program evaluation. As greater weight is placed on the results of the test, the links to and impact on instruction become stronger and stronger (p. 181).

There is always the concern as to whether achievement test results are ever used. Tests can be highly motivating for students especially if results are received quickly and with explanations that have some meaning to them (Ebel, 1981; Ahmann & Glock, 1981; Shertzer & Linden, 1979; & Linn, 1983).

The chairmen of a 1979 Conference sponsored by the National Institute of Educational Research on Testing (cited in Linn) commented

in their summary that "current testing procedures are not helpful to teachers or students in their day to day efforts to teach and learn" (p. 181). One of the Committees had stated that "present-day testing programs are largely extraneous to every day classroom teaching" (p. 181).

While test publishers' manuals state their purposes to be extremely useful in "every day classroom teaching", (Hieronymus, Lindquist & Hoover, 1982 & SRA, 1979 (cited in Linn); CAT Examiners Manual, 1981) it appears to be difficult to determine what is actually happening with test results.

Results of a study done by Stetz & Beck (1981) showed that in their sample of 3000 teachers, 59% indicated "some" or "considerable" personal use of standardized test results. Linn (1983) cautions, however that actual practise may vary from teachers opinions expressed and statements made.

Linn further cited a study by Stake & Easley (1978) which concluded "teachers did not appear to have much taste for the information tests could provide about individual student problems or problems with their own teaching" (p. 182). Stoke & Easley further quoted "studies by Hotvedt (1978), Hastings et al., (1960) and Scheyer (1977) "which found that teachers valued their own judgement more highly than information from tests which were found to be little used in making instructional decisions" (p. 182).

Linn (1983) concludes in his look at strengthening the links between testing and instruction that "test publishers and researchers

are placing increasing emphasis on specific content and performance and on the diagnosis of problems rather than only focusing on normative results and global scores" (p. 182).

Two achievement test batteries which have moved in that direction are the Canadian Achievement Tests (CAT) and the Canadian Tests of Basic Skills (CTBS). The CAT examiners manual outlines its rationale:

> The Canadian Achievement Tests are a series of test batteries that represent a new concept in achievement testing. They combine the most important and useful characteristics of norm-referenced and criterion-referenced assessment. This combination provides information about the relative ranking of an individual student against a norm group. It also provides specific information about the instructional needs of the student (p. 1).

The CTBS Teachers Guide (1982) outlines seven purposes which the tests are designed to serve. Three which relate to this discussion are:

- to diagnose specific qualitative strengths and weaknesses in a pupil's educational development
- to indicate the extent to which individual pupils have the specific readiness skills and abilities needed to begin instruction or to proceed to the next step in a planned instructional sequence
- to diagnose strengths and weaknesses in group performance (class, building, or system) which have implications for change in curriculum or instructional procedures or emphasis (p. 3).

In a critical review of the two tests Bachor & Summer (in press) comment:

> If the two tests are compared as norm-referenced batteries, CAT is technically adequate while the CTBS is of questionable adequacy. Looking at the two as criterion-referenced tests, neither one is satisfactory. Both lack sufficient items to place much confidence in the objective based analysis suggested (p. 28).

It would be of interest - against the backdrop of information on standardized aptitude and achievement measures to see how an aptitude and achievement test would compare with student end of year report card marks.

Carrying on in the search for effective decision making tools in the school, the area of teacher-made tests will be explored.

E. Teacher-Made Tests

As defined by Anastasi (1976), teacher-made tests belong at the specificity end of a continuum of tests of developed abilities.

FIGURE 2 Tests of Developed Abilities: Continuum of Experiential Specificity

Specificity ..... Generability				
Course-oriented achievement tests	Broadly oriented achievement tests	Verbal-type intelligence and aptitude tests	Non-language tests	Culture-fair tests

Note: From Psychological testing (p. 395) by A. Anastasi, 1982, New York: Collier. Copyright 1982 by Collier.

They are developed by the classroom teacher for use directly within the context of a particular unit of material and its related instructional objectives.

Gronlund (1982) adapted the following chart [SEE FIGURE 3] from Airasian & Madaus (1972), which outlines characteristics of four kinds of achievement tests which teachers construct. He points out the purposes may overlap at times and serve more than one function.

Four specific features of classroom tests - which are not exhaustive, are often lacking in standardized tests. According to Linn (1983) these are:

1. the degree of match between test items and instructional objectives
2. the use of test results to provide feedback to students and to the teacher
3. the use of tests to flag facts or concepts that are considered important
4. the use of tests to determine grades (p. 179).

This first point stresses the need for content validity. Not only must a teacher be concerned with "match" but also with selecting a balanced sample of subject-matter topics from the curriculum. "Thus, how adequately the test has sampled the intended outcomes is a vital question" (Gronlund, 1982, p. 127).

As much as it is recognized that standardized achievement tests can never replace teacher-made tests, they continue to come under constant

FIGURE 3 Characteristics of Four Types of Achievement Tests

TYPE OF TEST	FUNCTION OF TEST	SAMPLING CONSIDERATIONS	ITEM CHARACTERISTICS
	Measure prerequisite entry skills	Include each prerequisite entry behavior	Typically, items are easy and criterion-referenced
PLACEMENT	Determine entry performance on course objectives	Select representative sample of course objectives	Typically, items have a wide range of difficulty and are norm-referenced
FORMATIVE	Provide feedback to students and teacher on learning progress	Include all unit objectives, if possible (or those most essential)	Items match difficulty of unit objectives and are criterion-referenced
DIAGNOSTIC	Determine causes of recurring learning difficulties	Include sample of tasks based on common sources of learning error	Typically, items are easy and are used to pinpoint specific causes of error
SUMMATIVE	Assign grades, or certify mastery, at end of instruction	Select representative sample of course objectives	Typically, items have a wide range of difficulty and are norm-referenced

Note: Adapted from P.W. Airasian and G.F. Madaus, "Functional Types of Student Evaluation," Measurement and Evaluation in Guidance, 4(1972), 221-33). As cited in Gronlund (1982) p. 19.

scrutiny and criticism (Ebel, 1979, 1980; Ahmann & Glock, 1981; Macdonald, 1975).

Ebel (1979) summarizes many of these criticisms in his discussion of common weaknesses of teacher constructed tests:

(a) reliance on subjective judgements,

(b) reliance on absolute standards of judgement

(c) hasty test preparation

(d) use of short, inefficient tests

(e) testing trivia

(f) careless wording of questions

(g) neglect of sampling errors

(h) failure to analayze the quality of the test (p. 66).

Relevance and discriminating power are two factors which Ebel (1980) considers to be essential in a properly constructed teacher-made test. Utilizing important facts from the course objectives must be put before trivial details and overall course content. Understanding and application must be stressed rather than memorization of facts. Being able to select appropriate test items and structure them clearly will increase the discriminating power of the test and thus its reliability.

Reliability and validity of teacher-made tests is another contentious issue in the educational measurement field.

Hills (1981) advises that a "teacher should try to plan and give tests so that these measurements are reliable and should evaluate the reliability of these measurements routinely" (p. 46). Additionally, the teacher also has a heavy responsibility to ensure that school

constructed tests are clear, that they measure accurately and that the results are used validly (Ebel, 1979).

Linn (1983) concludes that neither teacher-made tests nor standardized achievement tests can do an effective job on their own. Both are needed to give a clear and composite view of students' academic progress.

Investigating teacher-made test results as transposed into various report card marks provide more information on this component of the evaluation process. Looking at how the four core areas connect and interplay may give us some clues as to the type of tests the teachers have constructed. In addition, it may also be interesting to compare the marks to a standardized achievement and aptitude test.

Alternate methods of assessing a class will not be explored.

F. Rank-Ordering of Students

Teachers avail themselves of numerous decision making processes to assess their students other than formally constructed tests. Rank-ordering of the class is one such method. It utilizes within group comparison of differences with respect to a set of specific criteria. Shertzer and Linden (1979) define rank-ordering as follows:

> The names of the members of a given group (set of objectives, ideas or events) are arranged in serial order from high to low in keeping with their status in terms of the characteristics being judged. The rank of 1 is assigned to the highest placed member, 2 to the next highest, and so on until the last member is ranked (p. 406).

Rank ordering as a system of rating students can be simple and straight forward but can also become unwieldy and difficult to use as the number of students and the number of criteria being assessed increase.

As Guilford (1954) points out "the use of ratings rests on the assumption that the human observer is a good instrument of quantitative observation, that he/she is capable of some degree of precision and some degree of objectivity" (p. 278).

The teacher utilizing any type of ordering must observe, record and evaluate student procedures and products. As Shertzer and Linden (1979) establish, this process can be fraught with inaccuracies. Errors could be found in:

1. Personal bias where the rater may be too lenient and either rate too high or too low on a scale.
2. The halo effect which may direct the rating according to the rater's general impression of the student.
3. Judgements being made that a student trait is either similar to or opposite to the raters.
4. Too many scores clustering around the midpoint of the scale.

Ahmann and Glock (1981) note that "ranking can be fairly reliable" (p. 184) but that "as the basis for the ranking broadens, its reliability tends to drop" (p. 185). This drop in reliability is influenced as the characteristics being rated lose their clear meaning and as the order of their importance becomes less definitive.

Reliability can be increased by the same person repeating the ranking or another person substantiating the results with a similar ranking.

Shertzer and Linden (1979) caution that ratings of any kind should not be used alone in making decisions about students. This and other data must be integrated to give a full picture of the situation.

One alternate use of this ranking method, is the practise of American colleges requiring a student's "rank-in-class" for admission purposes. The debate has raged since the 1960's about the dangers of class ranking but the practise continues (Newcomer, 1972; Cavalier, 1972; Kelleher, 1982).

This leaves us with the question of whether class rank-orderings add any useful information to our decision making quest. It would be of value to see how performance (as measured by final grades and standardized tests) rates with class rankings.

Arranging the collected information into meaningful grades becomes the next task.

G. Assigning Grades

For a classroom teacher, assigning grades or report card marks, is fraught with many dangers. Unless the teacher is extremely clear in defining what those marks mean, instructor bias can intrude and reduce the validity of these grades.

Ebel (1979) elaborates further:

> The lack of clearly defined, uniform bases for marking and standards for the meaning of various marks tends to allow biases to lower the validity of marks. Often such extraneous factors as

> pleasantness of manner, willingness to participate in class discussions, skill in expressing ideas orally or in writing, or success in building a self-image as an eagar capable student, will influence an instructor's assignment of grades (p. 233).

Teacher assigned grades should then be reflecting only the teacher's judgement of the quality of a student's performance in achieving instructional objectives (Terwilliger, 1977) and "not the amount of effort expended, the student's work habits, attitude, character traits or personality traits" (p. 22).

The teacher must also produce "sufficient, relevant, objective evidence to use as a basis for assigning marks" (Ebel, 1979, p. 233) or they tend to be unreliable.

According to Terwilliger (1977), a teacher must communicate several factors to the students:

> 1) the major objectives of the course, 2) the general criterion measures by which achievement of the course objectives is to be determined (tests, projects, assignments, etc.), and 3) the way in which data from the criterion measures are to be employed and/or weighted in arriving at course grades (p. 22).

Although there are many functions of grades (Hills, 1981) the primary purpose must be to communicate "about the degree of achievement of academic competence in a particular subject" (p. 287). Other functions such as those related to motivation and helping children to learn about life in a competitive society must be kept in perspective.

Hills suggestion that grades also keep teachers honest is a subtle reminder that grades cannot be used to punish a child.

Macdonald (1975) elaborates on this last point of Hills by reminding teachers of their moral obligations in consciously choosing a personal evaluation framework and being prepared to live with the consequences. He claims that this means serving the needs of the student not just the system.

Hotard (1983) looked at the relationship between standardized achievement tests and final teacher assigned grades.

FIGURE 4 8th Grade Achievement Scores with teacher grades

Reading r = .36
English r = .34
Math r = .43

> The correlations computed [SEE FIGURE 4] are significant $p<.01$ for 4th, 6th, 8th and 10th grades. These moderate correlations agree with previous research and reflect the fact that teacher grades are limited in reliability and validity, the 100 students in this sample were in many different schools and were probably assigned grades based on multiple criteria. The consistency of the relationships between teacher grades and achievement tests supports the concurrent validity of achievement tests while at the same time demonstrating that there is a great deal of unspecified variance between teacher grades and achievement test scores (p. 6).

Jubilee Junior High School in Edson provides an example of one school's grading system. When the four core subject areas were

organized into "official" departments at Jubilee, a uniform marking system was developed.

All math students, in grade 7, for example, regardless of who their instructor was, were graded in the same manner. Each report was structured as follows:

FIGURE 5 Grade 7 MATH MARK DISTRIBUTION

Unit tests and report card exams	50%
Quizzes and assignments	30%
Attitude	10%
Notebook	10%
	100%

3 Reports + final exam = FINAL GRADE

3 (25) + 25% = 100%

A debate that ensued as a result of this reorganization, was one related to grading of students' attitudes. In spite of the fact that there was a double marking system in place [SEE FIGURE 6], many staff felt they also wanted to include attitude in the academic section. As a result, most departments included "attitude" for up to 10% of the term mark in the academic grades.

FIGURE 6 POSSIBLE REPORT CARD MARKS, JUBILEE JUNIOR HIGH

Academic Marks	Effort Marks
A = 100-80%	E - excellent
B = 79-65	S - satisfactory
C = 64-50	U - unsatisfactory
D = 40-49	
X = 39 and below	

No suggestion of the proportion of A's, B's, C's, D's, or X's or restriction as to the number which could be allotted in each class was given to the teachers.

The last concept to be investigated, and one which plays a subtle and often "not discussed" role in educational judgements is that of teacher-expectations of students.

H. Teacher Expectations

There is a vast body of research related to teacher expectations of student outcomes. For the purposes of this study, only a few points will be dealt with.

Rosenthal & Jacobson's (1968) Pygmalion study involved teachers expectations being experimentally manipulated. "Students in early grades for whom high teacher expectation had been induced showed significant gains in total IQ and reasoning IQ when compared to other students in their school" (Cooper, 1979, p. 390).

The whole concept of a teacher knowing or believing a child to be bright and treating him or her as such created a self-fulfilling prophecy. This appeared to start out as halo effect "when certain things are known or believed about a pupil, other things about him/her true things or not, are implied (Rosenthal & Racobsen, 1968, p. 54).

Updated research on the Pygmalion effect (Cooper, 1979) "concludes that expectations influence performance, but they likely sustain it at a pre-existing level or allow latent differences in student performance to emerge rather than radically alter its course" (p. 392). Cooper goes on to propose a model "outlining the cognitive processes through which

teacher expectations can sustain a given level of achievement" (p. 389), which he says would include criteria for identifying expectation-effect-prone teachers.

Willis (1974) (cited in Brophy & Good, 1974) research with teachers of young children, found that "most teachers are capable of making a generally accurate assessment of a child's abilities and potential simply on the basis of observing him/her in the classroom for a few days" (p. 182)

Brophy & Good's review of studies done related to teacher expectations concluded with numerous implications for teachers and teaching. These included developing "appropriate expectations by carefully monitoring the day-to-day behavior of students" and by assessing "the appropriateness of their (teachers) behavior with regard to how it affects students performance" (p. 363).

Pedulla, Airasian and Madaus (1980) asked teachers in Ireland to rate students on IQ, math and English and 12 social and academic behaviors. Standardized IQ math and English were then compared with the 15 teacher ratings.

> The major implication of the results from this study is that even in the absence of standardized test-score results, teacher judgements of students' intelligence and mathematics and English attainment tap a dimension similar to that tapped by standardized tests, but are also interwined with academically related behaviors such as attention span and persistence" (p. 307).

It appears that teacher's judgements of academic performance are confounded with their judgements related to other related behaviors such as attention span and persistence. Pedulla et al argue that the separation of these two judgements cannot be done. As might be expected, standardized tests are not confounded with classroom behaviors.

As professionals, teachers are capable of making generally accurate judgements regarding their student's abilities. On the basis of these judgements, expectations are formed and decisions are made concerning academic development. Teachers must constantly be aware of their own expectations and judgements and be ready to evaluate their behavior in relation to how it effects their students progress.

I. Overview

Within the decision making process surrounding student evaluation, there are various interconnected elements. Exploration of some of those elements and their relationships may provide some useful information about that process.

If report card grades -- the result of teacher-made tests and other forms of assessment — were correlated both within and between reporting periods, what would be found about the links connecting those grades?

Rank-ordering has been completed by the classroom teacher. Would any significant relationship become evident if those rankings were compared with 1) final grades; 2) a standardized achievements test; or 3) an aptitude test?

How will year end report card marks fare against a standardized achievement test or an aptitude test?

What type of connection is found in comparing an achievement test like the CAT with an aptitude test like the DAT?

III. METHOD

A. Sample

In this study, the Grade 7,8 and 9 students at Jubilee Junior High School in Edson were the sample population. There were the following number of students in each grade:

Grade	Enrolment
7	134
8	191
9	167

During the 1981-82 school year, Jubilee was the main junior high school in Edson. The Grade 7 population came mainly from two elementary feeder schools within the town. However, a small number of students who had completed Grade 6 in the outlying areas of Robb and Peers were bused in.

Overall, bused students accounted for approximately 60% of the total school population. The remainder of the students were Edson residents, the majority of whom lived within walking distance of the school.

Pinegrove Elementary School, one of the feeder schools, retained the grade seven students who lived in the town and walked to school.

Edson, in 1981-82, was a west-central Alberta town of approximately 8,000 population. Industry within the community included oil and gas exploration and production, coal mining, forestry and logging and marginal farming and ranching.

B. School Achievement Marks

In the 1980-81 school year, language arts, math, science and social studies were formally structured into departments with one teacher responsible for over-seeing curriculum and organizational details. Course outlines were formalized with all of the teachers in each department in agreement with the concept of a coordinated curriculum. Common final exams in each of the four core areas were constructed. Evaluation methods were agreed upon with each teacher using a similar marking scheme to arrive at final grades. In 1982, common finals were written for the second year. All four exams were scheduled during the third week in June and conducted with the entire grade writing in the gymnasium. The exams were graded by the teacher who had instructed the class.

During the 1981-82 school year numerous test scores from the grade 7, 8 and 9 students were collected for the purpose of this study.

Report card marks in language arts, math, science and social studies were collected over the three reporting periods of:

November 1981

March 1982

June 1982

The November report card marks included grades from various exams and assignments from the September to early November time period. Student classwork and tests from mid November to March made up the second reporting period marks. Final June grades included the November mark, the March mark, accumulated assignments and exams from April to

June and the final exam. Each of these four sections were weighted 25%, giving a cumulative total of 100%.

For report card purposes, students were assigned grade letters. The letters with corresponding numerical percentages were:

A 100 - 80%

B 79 - 65%

C 64 - 50%

D 49 - 40%

X 39% and lower

For the data comparison purposes, the following numerical assignments were given to the letter grades:

A - 5

B - 4

C - 3

D - 2

X - 1

C. Teacher Rankings of Language Arts and Math

Prior to the final reporting period in June, before compilation of the final third term marks was started and before the final June exams were administered, teachers were asked to rank their students. Each language arts and math teacher was asked to rank the students in each individual class. Teachers were requested to assign to the student they felt would attain the highest overall yearly grade a 1, with the second highest student being given a 2 and so on up to the number of pupils in that particular class.

As a result, students were ranked as follows:

7A:	1 to 23	8A:	1 to 24	9A:	1 to 26
7B:	1 to 23	8B:	1 to 22	9B:	1 to 22
7C:	1 to 21	8C:	1 to 16	9C:	1 to 16
7D:	1 to 22	8D:	1 to 25	9D:	1 to 21
7E:	1 to 20	8E:	1 to 27	9E:	not ranked
7F:	1 to 21	8F:	1 to 26	9F:	1 to 20
		8G:	1 to 23	9G:	1 to 21
		8H:	1 to 23	9H:	1 to 17

Grade 7 language arts and math teachers completed rankings for all six classes. Language arts rankings were available for all grade 8 students. Math rankings were available for four of eight classes: 8A, 8C, 8D and 8E. Language arts rankings were available for four of eight grade 9 classes: 9A, 9C, 9D and 9F. Math rankings were received for six of eight grade 9 classes: 9A, 9B, 9C, 9D, 9G and 9H.

Only students for whom grades were available for all three reporting periods were included in this study. As a result, students whose marks were incomplete or who transferred in or out of the school during the 1981-82 school term were deleted in the collected data.

D. Canadian Achievement Tests

Canadian Achievement Tests (CAT) were completed in May 1982 by all grade seven, eight and nine students. Form A, levels 16, 17, 18 and 19 were administered.

The entire school wrote the CAT in their respective homerooms with directions given over the intercom by the guidance counsellors. Teacher advisors supervised their homeroom classes while the tests were written.

For this study, scale scores for each student were recorded. The following subtests were included:

Reading vocabulary	(RV)
Reading comprehension	(RC)
Reading total	(RT)
Spelling	(SP-C)
Language mechanics	(LM)
Language expression	(LX)
Language total	(LT)
Mathematics computation	(MCP)
Mathematics concepts & applications	(MC&A)
Mathematics total	(MT)
Total battery	(TOT)
Reference skills	(RS)

The testing was divided into five time frames as outlined in the CAT Examiner's Manual. Locator tests, which were used to determine which of the four test levels most suited the individual students level, were administered by teacher-advisors during an advisor session. The four testing periods were then scheduled in a staggered arrangement to avoid using the same period two days in a row.

CAT scheduling over a four day frame was:

DAY			
1	2	3	4
Reading Vocabulary Reading Comprehension	Spelling Language Mechanics Language Expression	Mathematics Computation Mathematics Concepts & Application	Reference Skills

The week before testing began, the counselling staff lead an inservice for the teaching staff on testing rationale, purposes and administrative details. Teacher-advisors were then able to outline to their advisory group the purpose of completing the CAT.

The Locator Tests were completed and scored by teacher-advisors. The appropriate level answer sheets and test booklets were distributed by the counsellors. All organizational preparation was completed prior to Day 1 of testing by teacher-advisors.

In a comparison of the CAT item content with the Alberta curriculum, it is this writer's opinion that the CAT has face validity and is consistent with the prescribed curriculum.

There were no arrangements made to complete the CAT with students who were ill or absent for any of the test sessions.

E. Differential Aptitude Test

Differential Aptitude Test (DAT) results were collected for grade 9 students. The DAT, form S, was written in February of 1982. There were no DAT scores available for grade 7 or 8 students.

For this study, DAT raw scores were used in reporting the following subtests:

Verbal reasoning	(VR)
Numerical ability	(NA)
VR & NA	(VR & NA)
Abstract reasoning	(AR)
Clerical speed & accuracy	(CL.SP.)
Mechanical reasoning	(MR)
Space relations	(SR)
Spelling	(SP)
Grammar	(GR)

Grade 9 students were prepared to write the DAT during their guidance classes. At that time, the guidance counsellors whose teaching responsibilities were split between the eight classes, provided standard information regarding the DAT. Explanation was given regarding the purpose of the tests and their future use. Students were advised they would receive the test results and be able to use them in career planning and for grade ten program placement.

Actual administration of the subtests were conducted over the school intercom system by one of the counsellors. Grade 9 teachers supervised the writing of the exams on a homeroom-advisor basis.

Students remained in their homeroom classroom for the entire testing period. They were allowed to bring free reading material with them for use if they finished the exam early. Teacher supervisors ensured students were quiet until the entire class was finished

writing. Short stretch breaks were given during Days 1 and 3 of testing. For Day 2, a ten minute break after the second subtest was allowed.

DAT scheduling over a three day time period was:

DAY		
1	2	3
Verbal Reasoning Numerical Ability	Abstract Reasoning Clerical Speed & Accuracy Mechanical Reasoning	Space Relations Spelling Grammar

The timetabling was staggered to facilitate using different class periods each day. Day 2 was scheduled during the afternoon while Days 1 and 3 occurred in the morning.

Students who missed any of the test days were tested the following week. They were removed from their regular classes and one of the counsellors completed the testing with them.

Test results and follow up interpretation was provided to the students during subsequent guidance classes.

F. Data Analysis

The following four measures were used in this study:

1) school achievement marks as reported in November, March and June report cards
2) teacher rank-orderings of students in language arts and math
3) Canadian Achievement Test results
4) Differential Aptitude Test results for Grade 9.

In an attempt to look at the relationship between all four measures, Pearson Product Moment Correlation Coefficients were obtained.

Spearman's rho could have been used when calculating correlations of school achievement marks and the rank-orderings of students. However, given the large sample in this study, the correlations obtained by the Pearson Product Moment should be consistent with the Spearman rho correlations.

IV. RESULTS

A. Grade Seven Students

Table 1 presents the number of subjects, means and standard deviations for three school reporting periods and the Canadian Achievement Tests (CAT). The three report card periods have means based upon school grades being given a rank of 5 for an A standing, 4 for a B, 3 equals a C, 2 a D and 1 is an X. The LA and MATH rank-ordering means are rankings of the class which range from 1 to 20-23 for Grade 7. The CAT results are recorded as scale scores with corresponding grade equivalents noted.

Pearson Correlation Coefficients were obtained for the teacher-made test results for the reporting periods of November 1981, March 1982 and June, 1982. Table 2 shows all correlations are above .72 and are significant at the .05 level. The November and March report card results are independent of other tests results. However, the June results are cumulative year's work and are therefore confounded variables. The same teacher taught both language arts and social studies which correlated .80 in November 1981, and .82 in March 1982. There were two different individuals teaching math and science which correlated .72 in November and .74 in March 1982.

Table 3 indicates the correlations between the November and March reporting period and the November and June reporting period. The results are all over .67 with the highest correlations showing on the diagonal. Again the June results are confounded due to the fact that they are averages over the year's work.

Before final exams were written and year end grades completed, math and language arts teachers were asked to rank order their students. These rankings, when correlated with June marks, ranged from -.76 to -.89 (Table 4). The correlations appear as negative numbers due to the ranking process. Language arts rankings correlated highly with all end of the year marks, although a small discrimination was seen between the LA ranking and science. Math rankings also correlated highly with the other subjects, however they appear to discriminate better than the LA rankings.

In Table 5, LA rankings are strongest with CAT math and overall test totals at -.66. Math rankings correlate with CAT math and overall test totals at -.80 and -.73.

Pearson Correlation Coefficients between the CAT and June core subjects are shown in Table 6. The CAT reading total correlates with the four core areas from .53 to .59 and does not appear to discriminate greatly between the four. Spelling correlations are all low. Highest correlations with the CAT language total is seen with Sci_3 at .64. Again little discrimination is shown between subjects. Math total in the CAT has the strongest relationship. It correlates .81 with $Math_3$ and .71 to .72 with the other three core area subjects. Here again little independence of grades is seen. Overall the CAT correlates in the .73 to .76 range with the four core subjects.

TABLE 1 Number of Subjects, Means and Standard Deviations, Grade 7

	NUMBER	MEAN LETTER GRADES	STANDARD DEVIATION
1. Report Card Marks			
LA_1 (Nov. '81)	127	3.4	1.1
$Math_1$	129	3.8	1.1
Sci_1	129	3.2	1.1
Soc_1	130	3.5	1.2
LA_2 (Mar '82)	131	3.3	1.3
$Math_2$	133	3.3	1.3
Sci_2	134	3.2	1.2
Soc_2	134	3.3	1.2
LA_3 (June '82)	131	3.3	1.2
$Math_3$	134	3.2	1.3
Sci_3	134	3.2	1.1
Soc_3	134	3.6	1.2

		MEAN CLASS RANK	
2. Rank-Ordering			
LA Rank	129	11.3	6.3
MATH Rank	132	11.6	6.5

	NUMBER	GRADE EQUIVALENT	MEAN SCALE SCORES	STANDARD DEVIATION
3. CAT				
Reading Vocabulary	131	7.8	527.2	62.2
Reading Comprehension	131	7.1	550.3	63.3
Reading Total	131	7.6	539.4	61.5
Spelling	130	7.6	538.2	74.3
Language Mechanics	130	6.7	552.8	77.6
Language Expression	130	6.4	539.4	59.9
Language Total	129	6.3	543.8	61.0
Mathematics Comprehension	129	7.1	489.6	53.5
Mathematics Concepts & Application	129	7.5	506.0	56.3
Mathematics Total	128	7.3	492.4	54.7
Test Total	128	7.2	509.4	56.3
Reference Skills	130	7.3	543.9	72.1

TABLE 2 Correlations Between Core Subject Areas, Three Reporting Periods, Grade 7

NOVEMBER 1981

	$MATH_1$	SCI_1	SOC_1
LA_1	.717	.725	.804
$MATH_1$		.723	.741
SCI_1			.724

$N = 127$

$p < .05$

MARCH 1982

	$MATH_2$	SCI_2	SOC_2
LA_2	.776	.792	.818
$MATH_2$		.735	.716
SCI_2			.753

$N = 130$

$p < .05$

JUNE 1982

	$MATH_3$	SCI_3	SOC_3
LA_3	.814	.843	.885
$MATH_3$		.843	.832
SCI_3			.854

$N = 131$

$p < .05$

TABLE 3 Correlations Between Core Subject Areas, Grade 7

	MARCH 1982			
NOVEMBER 1981	LA_2	$MATH_2$	SCI_2	SOC_2
LA_1	.828	.757	.726	.742
$MATH_1$	.710	.794	.757	.672
SCI_1	.756	.713	.793	.714
SOC_1	.796	.710	.745	.813

$N = 126$

$p < .05$

	JUNE 1982			
NOVEMBER 1981	LA_3	$MATH_3$	SCI_3	SOC_3
LA_1	.876	.768	.761	.806
$MATH_1$	.734	.878	.757	.755
SCI_1	.772	.760	.860	.769
SOC_1	.802	.760	.747	.862

$N = 126$

$p < .05$

	JUNE 1982			
MARCH 1982	LA_3	$MATH_3$	SCI_3	SOC_3
LA_2	.901	.777	.808	.821
$MATH_2$	.785	.871	.796	.774
SCI_2	.766	.820	.856	.796
SOC_2	.817	.746	.771	.857

$N = 130$

$p < .05$

TABLE 4 Teacher Rankings of Students in LA and MATH Correlated With Core Subject Achievement for June 1982, Grade 7

JUNE 1982

	LA_3	$MATH_3$	SCI_3	SOC_3
LA Rankings	-.858	-.808	-.767	-.821
MATH Rankings	-.794	-.891	-.761	-.780

$N = 126$

$p < .05$

TABLE 5 Correlations Between LA and MATH Rankings and Canadian Achievement Test Subtest Totals, Grade 7

	READING TOTAL	LANGUAGE TOTAL	MATHEMATICS TOTAL	CAT TOTAL
LA Rankings	-.469	-.530	-.664	-.657
MATH Rankings	-.534	-.535	-.800	-.730

$N = 126$

$p < .05$

TABLE 6 Correlations Between the Canadian Achievement Tests and Core Subject Areas, June 1982, Final Report Period, Grade 7

CANADIAN ACHIEVEMENT TESTS	JUNE 1982			
	LA_3	$MATH_3$	SCI_3	SOC_3
Reading Vocabulary (RV)	.480	.462	.517	.456
Reading Comprehension (RC)	.562	.534	.582	.571
Reading Total (RT)	.563	.533	.588	.555
Spelling (SP-C)	.345	.339	.396	.372
Language Mechanics (LM)	.604	.567	.597	.586
Language Expression (LX)	.473	.460	.530	.474
Language Total (LT)	.604	.571	.637	.594
Mathematics Computation (MCP)	.705	.765	.670	.694
Mathematics Concepts & Application (MC&A)	.587	.741	.651	.627
Mathematics Total (MT)	.705	.814	.713	.715
Test Total (TOT)	.743	.751	.761	.734
Reference Skills (RS)	.509	.528	.565	.609

$N = 125$

$p < .05$

B. Grade Eight

Table 8 presents the number of subjects, means and standard deviations for three school reportings periods and the Canadian Achievement Tests (CAT). The report card period means are based upon school grades being ranked as: 5 for an A, 4 for a B standing, 3 equals a C, 2 a D and 1 is an X. The LA and MATH rank-ordering means are rankings of the class which range from 1 to 16-27 for Grade 8. The CAT results are recorded as scale scores with corresponding grade equivalents noted.

Pearson Correlation Coefficients were obtained for the teacher-made test results for the reporting periods of November 1981, March 1982 and June 1982. Table 8 indicates all correlations are above .65. November and March results are independent of other tests results. However, the June results are cumulative years work and are therefore confounded variables.

In the grade eight classes subjects were fully departmentalized. As a result, none of the core teachers taught more than one subject. There were no grade eight classes having the same teacher instructing them in language arts and social studies or math and science. Correlations between language arts and social studies in November were .74 and in March .68. In math and science correlations in November were .76 and in March .73.

Table 9 indicates the Pearson Correlation Coefficients between the November and March reporting period and the November and June reporting period. Correlations are all over .61 and are significant at the .05

level with the highest correlations showing on the diagonal. The June results are confounded as they include averages over the years work.

When LA rankings and math rankings were correlated with the June results, these rankings ranged from -.54 to -.79. Table 10 outlines the results. The grade 8 rank-order correlations were not as high as the grade 7 correlations. However, the same pattern was evident, with LA rankings correlating highly with LA_3. Some discrimination was evident in the LA rankings with Sci_3 and to a smaller degree with Soc_3. The math rankings correlated highly with $Math_3$. It discriminated the best with Soc_3. In Table 11, correlations between LA rankings and the CAT are moderate. However, little discrimination is shown by the rankings between the CAT subtests. A similar phenomena is seen with the Math rankings.

Pearson Correlation Coefficients between the CAT and June core subjects are shown in Table 12. The CAT reading total correlates highest with Soc_3 at .55. Sci_3 correlates higher than LA_3 at .53 to .47. Spelling correlates higher than either grade 7 or 9 results indicate. However, relationships are moderate in the .41 to .45 range. CAT language subtest total has a moderate relationship in the .52 to .58 range with all four core areas. However, little discrimination is shown between academic subjects. The CAT math total correlates strongly with $Math_3$ at .71. Here some independence of relationship is seen as LA_3 correlates less well. The CAT test total indicates $Math_3$ correlates highest at .67 with Sci_3, Soc_3 and LA_3 following in order to .56.

TABLE 7 Number of Subjects, Means and Standard Deviations, Grade 8

	NUMBER	MEAN LETTER GRADES	STANDARD DEVIATION
1. Report Card Marks			
LA_1 (Nov. '81)	190	3.6	1.3
$Math_1$	190	3.0	1.4
Sci_1	190	3.3	1.2
Soc_1	191	3.4	1.3
LA_2 (Mar '82)	190	3.5	1.2
$Math_2$	190	2.9	1.4
Sci_2	190	3.5	1.2
Soc_2	191	3.7	1.1
LA_3 (June '82)	188	3.3	1.2
$Math_3$	189	2.7	1.4
Sci_3	189	3.2	1.2
Soc_3	189	3.5	1.2

		MEAN CLASS RANK	
2. Rank-Ordering			
LA Rank	181	12.0	7.0
MATH Rank	92	12.3	7.2

	NUMBER	GRADE EQUIVALENT	MEAN SCALE SCORES	STANDARD DEVIATION
3. CAT				
Reading Vocabulary	188	8.8	557.8	76.7
Reading Comprehension	188	9.6	582.6	69.7
Reading Total	188	9.0	571.5	69.3
Spelling	187	8.1	552.8	78.6
Language Mechanics	188	9.0	579.5	80.2
Language Expression	188	8.3	568.4	64.1
Language Total	188	8.5	572.3	65.0
Mathematics Comprehension	188	8.2	512.5	65.0
Mathematics Concepts & Application	185	8.5	537.6	71.1
Mathematics Total	185	8.3	524.7	68.1
Test Total	184	8.5	543.8	67.7
Reference Skills	185	9.2	573.6	79.1

TABLE 8 Correlation Between Core Subject Areas, Three Reporting Periods, Grade 8

NOVEMBER 1981

	$MATH_1$	SCI_1	SOC_1
LA_1	.665	.733	.738
$MATH_1$		.756	.701
SCI_1			.753

$N = 189$

$p < .05$

MARCH 1982

	$MATH_2$	SCI_2	SOC_2
LA_2	.651	.658	.683
$MATH_2$		.729	.658
SCI_2			.752

$N = 190$

$p < .05$

JUNE 1982

	$MATH_3$	SCI_3	SOC_3
LA_3	.704	.728	.768
$MATH_3$		.777	.743
SCI_3			.800

$N = 188$

$p < .05$

TABLE 9 Correlation Between Core Subject Areas, Grade 8

MARCH 1982

NOVEMBER 1981	LA_2	$MATH_2$	SCI_2	SOC_2
LA_1	.785	.667	.682	.722
$MATH_1$	.610	.812	.724	.667
SCI_1	.673	.707	.778	.756
SOC_1	.715	.647	.684	.780

$N = 189$

$p < .05$

JUNE 1982

NOVEMBER 1981	LA_3	$MATH_3$	SCI_3	SOC_3
LA_1	.856	.672	.693	.752
$MATH_1$	.639	.891	.737	.698
SCI_1	.715	.771	.862	.769
SOC_1	.732	.684	.702	.859

$N = 187$

$p < .05$

JUNE 1982

MARCH 1982	LA_3	$MATH_3$	SCI_3	SOC_3
LA_2	.842	.649	.648	.733
$MATH_2$	.671	.881	.707	.675
SCI_2	.690	.756	.852	.746
SOC_2	.723	.702	.766	.904

$N = 188$

$p < .05$

TABLE 10 Teacher Rankings of Students in LA and MATH Correlated with Core Subject Achievement for June 1982, Grade 8

JUNE 1982

	LA_3	$MATH_3$	SCI_3	SOC_3
LA Rankings	-.768	-.723	-.612	-.674
N = 180				
MATH Rankings	-.696	-.789	-.637	-.544
N = 92				
$p < .05$				

TABLE 11 Correlations Between LA and MATH Rankings and CAT Subtest Totals, Grade 8

JUNE 1982

	Reading Total	Language Total	Math Total	CAT Total
LA Rankings	-.324	-.415	-.393	-.431
N = 176				
MATH Rankings	-.355	-.490	-.474	-.491
N = 91				
$p < .05$				

TABLE 12 Correlations Between the Canadian Achievement Tests and Core Subject Areas, June 1982 Final Report Period, Grade 8

CANADIAN ACHIEVEMENT TESTS	JUNE 1982			
	LA_3	$MATH_3$	SCI_3	SOC_3
Reading Vocabulary (RV)	.404	.419	.484	.513
Reading Comprehension (RC)	.437	.398	.490	.496
Reading Total (RT)	.456	.443	.531	.547
Spelling (SP-C)	.407	.423	.454	.417
Language Mechanics (LM)	.516	.536	.477	.524
Language Expression (LX)	.413	.444	.497	.442
Language Total (LT)	.524	.557	.553	.545
Mathematics Computation (MCP)	.520	.686	.562	.545
Mathematics Concepts & Application (MC&A)	.410	.653	.540	.473
Mathematics Total (MT)	.482	.706	.584	.534
Test Total (TOT)	.557	.670	.644	.620
Reference Skills (RS)	.409	.456	.463	.404

$N = 183$

$p < .05$

C. Grade Nine Students

Table 14 presents the number of subjects, means and standard deviations for three school reporting periods, the Canadian Achievement Tests (CAT) and the Differential Aptitude Tests (DAT). The CAT results are recorded as scale scores. The report card periods all have means based upon the letter grades being ranked as: 5 for an A standing, 4 for a B, 3 equals a C, 2 a D and 1 is representative of an X. The LA and MATH rank-ordered means are rankings of the class which range from 1 to 16-26 for Grade 9. The CAT results are recorded as scale scores with corresponding grade equivalents noted. The DAT results are recorded as raw scores.

The grade equivalent scores on the CAT for the three grades show an interesting pattern. Although Grade 7 scores are somewhat low, by Grade 8 the scores are becoming stronger especially in the language mechanics and expression subtests. By Grade 9 the student grade equivalents have improved subtantially, indicating the students are leaving the junior high school with a strong skill base as measured by the CAT.

Pearson Correlation Coefficients were obtained for the teacher-made test results for the reporting periods of November 1981, March 1982, and June 1982. Table 14 indicates all correlations range from .59 to .73. November and March results are independent of other test results. However, the June results are cumulative years work and are therefore confounded variables.

In grade nine, in most instances, a class of students had the same teacher for both language arts and social studies. In science and math

most classes also had a common teacher. As a result, grade nine students had 2 core teachers only, not 4 as in grade 8 or 3 as in grade 7. In November, language arts and social studies correlated .70 and in March .64, while math and science correlated .67 in November and .73 in March.

Table 15 shows the Pearson Correlation Coefficients between the reporting periods of: November and March, November and June, and March and June. Correlations between November and March range from .49 to .78 and are the highest on the diagonal. Between November and June correlations fall between .51 and .83. Again the highest are on the diagonal. June results are confounded due to the fact that they are averages over the years work.

Table 16 outlines the results of the LA and Math rankings. The grade 9 rankings follow a similar pattern to both grade 7 and 8, however, correlations tend to be lower overall. LA rankings correlated moderately with LA_3 at -.58. Discrimination among the other subjects was most evident with Soc_3. Math rankings correlated higher with $Math_3$ at -.78. Some independence of grades was seen with LA_3 and Soc_3.

In Table 17, minimal correlation is seen between LA rankings and the CAT subtest total. No independence of marks is evident. However, the Math rankings indicate a moderate correlation with the CAT math total at -.52 and CAT test total at -.45. The math rankings appear to show a reasonable discrimination between subtests.

Table 18 shows Pearson Correlation Coefficients between LA rankings and the DAT scores to be minimal. There is no discrimination evident

between the VR and NA subtests. However, Math rankings show a moderate -.56 correlation with NA and -.50 with VR & NA with some discrimination shown with VR.

Pearson Correlation Coefficients between the CAT and June core subjects are shown in Table 19. The CAT reading total correlates highest with Soc_3 at .58. Sci_3 correlates higher than LA_3 at .47 and .46. Spelling correlations are all minimal. The language total of the CAT shows its highest relationship with Soc_3 and Sci_3. All scores range from .43 to .53 with little discrimination shown. CAT math total correlate highly with the $Math_3$ June marks at .69. Sci_3 shows some relationship at .61 while Soc_3 and LA_3 are lower in the .52 to .55 range. The CAT test total correlations indicate Soc_3 and Sci_3 relate highest at .64 and .62. $Math_3$ and LA_3 are lower at .58 and .57.

Pearson Correlation Coefficients between the DAT and June core subjects are shown in Table 20. LA_3 correlates moderately with VR and SP at .41 for both. The LA_3 relationship with GR correlated at .51, however, Soc_3 showed a stronger relationship than LA_3 in both VR and GR at .52 and .53. $Math_3$ correlated with the NA at .65. There is still a moderate relationship seen with LA_3 at .50, Sci_3 at .58 and Soc_3 at .53. The VR & NA combination correlates moderately with all four final grades ranging from .51 to .60. Little discrimination between subject areas is seen.

Table 21 shows the Pearson Correlation Coefficients for the CAT and the DAT. Correlations between the CAT reading subtests and the DAT verbal reasoning (VR) and grammar subtests (GR) are relatively strong

(.63 to .72). Spelling subtests between both tests are moderately high (.77). CAT spelling and DAT verbal reasoning and grammar correlate at .57. The language subtests of the CAT and the DAT verbal reasoning, spelling and grammar fall in the .62 to .71 range. Math totals in both tests follow a similar pattern at .79. The strongest correlation between the CAT and DAT is found between the CAT test total and the DAT, VR&NA at. .82.

Tables 22 and 23 summarize the correlations of the two standardized test results and the final grades. Generally, the DAT and CAT correlations are higher with each other than with the final teacher-assigned grades.

TABLE 13 Number of Subjects, Means and Standard Deviations, Grade 9

	NUMBER		MEAN LETTER SCORE	STANDARD DEVIATION
1. Report Card Marks				
LA_1 (Nov. '81)	166		3.5	1.2
$Math_1$	167		3.0	1.2
Sci_1	166		3.2	1.3
Soc_1	165		3.2	1.2
LA_2 (Mar '82)	167		3.2	1.1
$Math_2$	167		3.3	1.3
Sci_2	167		3.3	1.1
Soc_2	167		3.4	1.2
LA_3 (June '82)	167		3.4	1.0
$Math_3$	167		3.3	1.2
Sci_3	167		3.3	1.0
Soc_3	167		3.3	1.2
			MEAN CLASS RANK	STANDARD DEVIATION
2. Rank-Orderings				
LA Rank	84		11.3	6.5
Math Rank	123		10.8	6.2
		GRADE EQUIVALENT	MEAN SCALE SCORES	STANDARD DEVIATION
3. CAT				
Reading Vocabulary	167	10.2	583.9	81.4
Reading Comprehension	167	10.8	599.8	74.1
Reading Total	167	10.5	593.5	74.6
Spelling	166	9.5	579.8	84.5
Language Mechanics	165	11.1	604.2	79.0
Language Expression	164	10.5	590.8	67.7
Language Total	164	10.6	597.2	66.3
Mathematics Comprehension	165	9.3	563.8	76.0
Mathematics Concepts & Application	165	10.3	588.8	71.4
Mathematics Total	164	9.6	575.1	74.3
Test Total	163	9.9	582.8	71.4
Reference Skills	165	10.9	603.5	76.7

Table 13 cont.

4. DAT	NUMBER	MEAN RAW SCORES	STANDARD DEVIATION
Verbal Reasoning (VR)	165	21.6	9.7
Numerical Ability (NA)	165	20.9	6.8
VR & NA	165	42.4	14.7
Abstract Reasoning	166	35.4	8.6
Clerical Speed & Accuracy	166	44.6	12.0
Mechanical Reasoning	166	44.5	9.7
Space Relations	165	33.4	10.6
Spelling	165	65.9	16.3
Language Usage	163	28.9	8.9

TABLE 14 Correlations Between Core Subject Areas, Three Reporting Periods, Grade 9

NOVEMBER 1981

	$MATH_1$	SCI_1	SOC_1
LA_1	.621	.671	.701
$MATH_1$		.670	.590
SCI_1			.629

$N = 164$

$p < .05$

MARCH 1982

	$MATH_2$	SCI_2	SOC_2
LA_2	.622	.662	.640
$MATH_2$		.730	.590
SCI_2			.635

$N = 167$

$p < .05$

JUNE 1982

	$MATH_3$	SCI_3	SOC_3
LA_3	.673	.739	.795
$MATH_3$		.746	.660
SCI_3			.752

$N = 167$

$p < .05$

TABLE 15 Correlations Between Core Subject Areas, Grade 9

MARCH 1982

NOVEMBER 1981	LA_2	$MATH_2$	SCI_2	SOC_2
LA_1	.676	.550	.577	.641
$MATH_1$	.604	.776	.682	.630
SCI_1	.648	.634	.710	.653
SOC_1	.659	.485	.579	.700

$N = 165$

$p < .05$

JUNE 1982

NOVEMBER 1981	LA_3	$MATH_3$	SCI_3	SOC_3
LA_1	.791	.608	.640	.755
$MATH_1$	.606	.831	.702	.644
SCI_1	.659	.649	.792	.669
SOC_1	.665	.513	.631	.792

$N = 165$

$p < .05$

JUNE 1982

MARCH 1982	LA_3	$MATH_3$	SCI_3	SOC_3
LA_2	.804	.618	.704	.675
$MATH_2$	.659	.921	.731	.616
SCI_2	.699	.736	.876	.669
SOC_2	.702	.624	.682	.868

$N = 167$

$p < .05$

TABLE 16 Teacher Rankings of Students in LA and MATH Correlated with Core Subject Achievement for June 1982, Grade 9

JUNE 1982

	LA_3	$MATH_3$	SCI_3	SOC_3
LA Rankings	-.582	-.489	-.546	-.441
$N = 84$				
MATH Rankings	-.589	-.779	-.724	-.527
$N = 123$				
$p < .05$				

TABLE 17 Correlations Between LA and MATH Rankings and CAT Subtest Totals, Grade 9

JUNE 1982

	Reading Total	Language Total	Math Total	CAT Total
LA Rankings	-.204	-.226	-.257	-.276
N = 83				
MATH Rankings	-.287	-.363	-.520	-.450
N = 121				
$p < .05$				

TABLE 18 Correlations Between LA and MATH Rankings and DAT Scores, Grade 9

	VR	NA	VR&NA	AR	CL	MR	SR	SP	GR
LA Rank	-.299	-.295	-.333	-.192	-.112	-.137	-.031	-.322	-.254
N = 84									
MATH Rank	-.368	-.562	-.501	-.447	-.116	-.287	-.304	-.358	-.329
N = 119									

$p < .05$

TABLE 19 Correlations Between the Canadian Achievement Tests and Core Subject Areas, June 1982 Final Report Period, Grade 9

CANADIAN ACHIEVEMENT TESTS	JUNE 1982 LA3	MATH3	SCI3	SOC3
Reading Vocabulary (RV)	.384	.296	.385	.506
Reading Comprehension (RC)	.453	.328	.483	.555
Reading Total (RT)	.457	.341	.473	.576
Spelling (SP-C)	.350	.308	.375	.371
Language Mechanics (LM)	.495	.451	.525	.481
Language Expression (LX)	.354	.310	.365	.445
Language Total (LT)	.483	.429	.500	.531
Mathematics Computation (MCP)	.561	.725	.590	.559
Mathematics Concepts & Application (MC&A)	.410	.564	.550	.478
Mathematics Total (MT)	.521	.687	.605	.552
Test Total (TOT)	.569	.581	.618	.641
Reference Skills (RS)	.406	.417	.493	.447

N = 163

$p < .05$

TABLE 20 Correlations Between the Differential Aptitude Tests and Core Subject Areas, June 1982 Final Report Period, Grade 9

	JUNE 1982			
DAT	LA_3	$MATH_3$	SCI_3	SOC_3
Verbal Reasoning (VR)	.410	.357	.493	.523
Numerical Ability (NA)	.501	.646	.580	.528
VR & NA	.505	.538	.598	.594
Abstract Reasoning (AR)	.397	.455	.484	.465
Clerical Speed & Accuracy (CL)	.283	.299	.245	.258
Mechanical Reasoning (MR)	.137	.190	.291	.247
Space Relations (SR)	.190	.260	.295	.269
Spelling (SP)	.413	.386	.437	.382
Grammar (GR)	.513	.409	.555	.529

$N = 163$

$p < .05$

TABLE 21 Correlations Between Canadian Achievement Tests and Differential Aptitude Tests, Grade 9

	DAT								
CAT	VR	NA	VR&NA	AR	CL	MR	SR	SP	GR
Reading Vocabulary (RV)	.634	.328	.573	.374	.093	.302	.253	.602	.628
Reading Comprehension (RC)	.665	.419	.636	.501	.143	.292	.386	.601	.700
Reading Total (RT)	.702	.411	.657	.481	.135	.323	.340	.655	.722
Spelling (SP-C)	.568	.351	.541	.356	.216	.178	.215	.765	.583
Language Mechanics (LM)	.543	.519	.602	.483	.102	.222	.332	.550	.642
Language Expression (LX)	.643	.463	.642	.415	.043	.278	.268	.511	.591
Language Total (LT)	.695	.578	.730	.511	.073	.295	.342	.617	.709
Mathematics Computation (MCP)	.429	.753	.636	.671	.240	.257	.416	.405	.515
Mathematics Concepts & Applications (MC&A)	.549	.733	.705	.669	.199	.422	.509	.460	.546
Mathematics Total (MT)	.514	.787	.708	.712	.263	.353	.490	.452	.560
Test Total (TOT)	.743	.705	.821	.674	.194	.369	.460	.697	.768
Reference Skills (RS)	.525	.489	.575	.606	.200	.222	.490	.515	.601

$N = 160$

$p < .05$

TABLE 22 Like Subject Correlations, Math, Grade 9

	CAT MATH TOTAL	DAT NA
$Math_3$	.69	.65
CAT Math Total		.79

TABLE 23 Like Subject Correlations, Language Arts, Grade 9

	LA_3	DAT VR	DAT SP	DAT GR
LA_3		.410	.413	.513
CAT Reading Total	.457	.702	.655	.722
CAT Spelling	.350	.568	.765	.583
CAT Language Total	.483	.695	.617	.709

V. DISCUSSION

A. The Research Questions

The present study was conducted to investigate the relationship between teacher-made evaluations and standardized tests: two of the many components in the decision making process surrounding student progress. Acknowledging the fact that the school is a complex network of interrelated elements, it was felt that an investigation related to student evaluation would provide an excellent "sample of everyday life" which in itself would likely provide an invaluable insight into the "whole" of Jubilee Junior High School.

Four basic questions have been asked.

QUESTION ONE: What kind of relationship exists between grades based upon teacher-made tests and the independence of those grades among the four core subjects?

Speculation was that math-science correlations and language arts-social studies correlations for each of the three reporting periods would be high. It was further anticipated that the results of the two basic areas would be independent of each other and generally not show a connection.

The expected pattern of same subject and similar area did emerge in all three grades. However, no independence of marks between the two major fields was shown (TABLES 2, 8, & 14).

Core subject correlations for the November-March and the November-June comparisons (TABLES 3, 9, & 15) showed high correlations in a similar configuration to first "among test" comparisons. It would

normally be expected that the correlations would be lower as the course content changed between report card periods. Some variability in performance between subject areas was anticipated but not seen.

The same teacher taught language arts (LA) and social studies (SOC) to the respective classes in grades 7 and 9. This might account for some of the strength of the correlations except for the fact that grade 8 students had different individuals teaching those two subjects and the correlations there were still high. In math and science, grade 9 was the only grade where one teacher taught these two courses to one class. The pattern of high correlations was similar in all three grades in these subjects.

Confounded variables (averaging in earlier term work), in the third reporting period may also partially explain these high correlations.

The teacher final grades reflect no restrictions as to the number of A's, B's, C's, D's or X's in each of their classes. The resulting final grade means for each of the three grades are relatively high (TABLES 1, 7 & 13). With this lower variability in scores, correlations between subject areas are lower than would normally be seen if there were more intervals in the scale such as percent grades.

The high correlations lead to the suspicion - although it would be difficult to measure, that some other factor, perhaps related to teacher attitudes toward the students is being indicated. The research has shown that teacher's judgement of students academic performance is contaminated with their additional judgement of academic behaviors (Pedulla et al., 1980). This may be reflected in the consistent

students grades and the resulting high correlations on teacher-made tests.

Teachers had agreed to use common marking systems in their respective departments. Marks for attitude comprised usually 10% of each reporting period. Terwilliger (1977) pointed out that this practise does not adhere to the guidelines of assigning grades on the basis of academic achievement.

Summarizing question one, very little independence among the grades of the four core subjects is seen although correlations among the four are high. These results lead the writer to challenge the teachers' ability to construct tests which can discriminate between content areas.

QUESTION TWO: Does a link exist between teacher rank-ordering of students and 1) their final grades, 2) the CAT, and 3) the DAT?

It was speculated that the rankings and the corresponding course would correlate highly and that the other subjects would show independence of marks.

The LA rankings in grade 7 and 8 were strongly correlated with LA_3, however, little discrimination with the other subjects is shown. Grade 7 students had the same teacher for LA and SOC which may account for the strength of the correlations.

The MATH rankings in these grades appear to discriminate somewhat better. The math teachers only taught the students math - as a result they saw the students half as often as the LA-SOC teachers did. If there is teacher bias in the MATH rankings it may not be so pronounced as the LA rankings.

The grade 9 rankings show a similar pattern as the grade 7 and 8, but the overall correlation is lower.

The grade 9 LA rankings that were available for this study included only four of eight classes. The four classes that were ranked likely incorporated the four lowest classes academically. When the overall class ability does not show a wide range of variability, it may be that it is more difficult for teachers to rank their students.

Grade 9 was the only grade where the same teacher instructed both math and science to the same class. The Math rankings did discriminate between the two major areas of math-science and language arts-social studies.

Correlations between the LA rankings and the language and reading subjects of the CAT, do not show any consistent pattern. Here validity of the rankings must be questioned. They appear to be measuring the teacher's perceived ability of the student in the subject area but are probably confounded with other teacher judgement and experiences not related to pure achievement.

The MATH rankings show a strong correlation in grade 7 with the CAT math scores. In grade 8 the connection is not a discriminating one. This may be due to only half the math rankings being available in grade 8. Grade 9 results follow a pattern similar to grade 7 but are considerably lower.

Overall, the MATH rankings appear to give the most substantial results. They may tell us that it is easier to rank students in math than in language arts where scoring students becomes more subjective.

LA and MATH rank-orderings correlated with the DAT appear to fall into a similar pattern as was seen with the CAT. The MATH ranking shows a respectable relationship with NA and as a result also with VR + NA. Small correlations between the LA rankings and VR, SP and GR subtests on the DAT suggest that the rankings and the subtests are measuring two very different domains. If the CAT is used as a standard by which to measure academic achievement then teacher rankings in LA are measuring some other unknown dimension.

This would lead to the possibility that the rank-ordering process, especially as related to language arts, likely contains errors related to personal bias, the halo effect and similar - contrast trait judgements. Also, as staff members did exchange information about students both in the staffroom and in general hallway discussion, independence of teacher judgement must be questioned (Shertzer & Linden, 1983).

If class rankings are in some cases confounded with teacher bias, it may be related to style differences among teachers. Some of the teachers at Jubilee may be what Cooper (1979) called "expectation-effect prone" (p. 392).

The link that exists between the rank-ordering of the students and their final grades, CAT scores and DAT scores appears to be of little value. These rankings do not provide any worthwhile information to aid in making educational decisions. Teacher judgement becomes suspect and speculation can be made about non-academic biases interfering with that judgement.

QUESTION THREE: How will student year end grades correlate with the CAT and the DAT?

Initial perusual seems to indicate strong correlations in appropriate patterns in the relationship between the CAT and the year-end report card marks. However, the "patterns" that emerge are ones which lead to serious questions about the construction of the teacher-made tests.

The CAT reading total score correlates consistently with all four core areas, particularly in grade 7 and 8. This would lead us to believe that rather than measuring instructional objectives appropriately, teachers' tests have become "reading tests".

None of the core areas show substantial independence from each other. $Math_3$ and the CAT Math Total show the most independence especially in the grade 8 results. However, $Math_3$ correlations with the other subtests show consistency with the three remaining core subjects. Content validity of the teacher-made tests must be questioned.

Table 12, with Grade 8 results, shows an interesting phenomena, which is not apparent in either Grade 7 or 9. the math final grade is as predictive of language skills (as measured by the CAT) as the language arts final grade (.557 to .524). From these results, two questions must be asked;

1) What is being taught in language arts and in math classes? Is the course content covering the Alberta curriculum?
2) What is being tested in the teacher-made tests in these two areas?

A similar pattern exists in the CAT spelling subtest correlations with the four final grades.

Teacher-made tests must have a high degree of match between instructional objectives and test items (Linn, 1983). The tests must strive to be criterion-referenced, where appropriate, and not simply measure general learning (Ebel, 1979). While this writer has not reviewed the teacher-made tests these students sat for, one could presume from the correlated nature of the results that these teacher-made tests appear to be measuring general learning and are not specific enough to be measuring the appropriate course content. The tests may also have been inadvertently structured to measure a students' test wiseness or how well that student can play the "testing game". If the teacher-made tests are not measuring what they claim to be, then the resulting grades are suspect.

Curriculum match of the CAT with the Alberta curriculum is appropriate. The pattern of the strength of the correlation coefficients between the CAT and the core subject area final grades decrease as we move from grade 7 to grade 9. This may be the result of the CAT drifting away from the curriculum as the grade level increases, or it may be related to the emphasis the teacher is putting on various parts of the curriculum as the grade 9 students are prepared for high school. Grade 9 appears to be the point at which the subject instead of the student begins to take precedence in the classroom.

Since the four core grades show no strong independence among each other, they must be measuring a common element which may neither be

academic nor content oriented. If the correlations in the core subjects were able to discriminate, they would be much lower and there would be a much broader spread. On this basis we can conclude these teacher-made tests are inferior to standardized achievement tests.

The grade 9 final grades correlate with the DAT consistently, however neither as strongly nor as clearly as anticipated. Like the CAT correlations, little clear discrimination between subject areas is seen.

From question three it is apparent that although correlations between student year end grades and the CAT and DAT are consistent they are lower than was anticipated.

QUESTION FOUR: What is the relationship between the CAT and the DAT?

Since achievement and aptitude measures overlap to some degree, it is anticipated that the CAT and DAT should show a relatively high relationship. In this instance, the correlation matrices between the CAT and DAT are not unexpected. Both correlate strongly in common areas such as the CAT reading and language totals and the DAT verbal reasoning subtests; the two spelling areas; the CAT language total and the DAT grammar sections; the CAT math total and DAT numerical ability; the CAT test total and the DAT, VR&NA subtest combination. The DAT and CAT correlate less consistently with teacher-made tests as represented by the year end grades.

The two standardized tests affirm one another as anticipated. If the CAT is considered as a standard by which to look at the effectiveness of teacher-made tests, questions would need to be asked

concerning exactly what domains and objectives the teacher-made tests are measuring. Until this is clarified, the CAT may be more valuable in educational planning than the classroom tests. However, teacher-made tests are meant to deal with day to day curriculum measures and if constructed properly should be measuring those concerns.

B. Recommendations

The standardized testing program at Jubilee in 1981-82 made good use of the aptitude and interest end of the program. However, achievement test results from the CAT were not put to their maximum use, likely as the result of several factors.

The year before, when the decision was made to take part in the CAT norming study, it was a counselling and administration decision. Commitment from the teaching staff was not sought. It would enhance the standardized testing program for goals to be clearly established with the support of the teachers. If the goals of the program included being able to evaluate student progress against other students across the province and country, as well as, improving instruction and curriculum within individual classrooms and content areas, then procedures could be implemented to carry out these goals. The newly organized departments would have the CAT results to help in assessing both student progress and in evaluating whether instructional objectives are being met.

Use of the CAT results might also be optimized by setting the administration date for the tests to fall schedule. With results available at that time, students strengths and weaknesses could be noted

and appropriate curriculum changes initiated. Further diagnostic testing may also be deemed necessary as a result of some test scores.

Tests results from the CAT could also be shared with both students and parents during the fall parent-teacher interview. These type of test results can often provide a motivational impetus for students in that they know where they stand in relation to the rest of the class. Students would also be informed - in an alternate method than report cards, where their strengths and weaknesses lie.

Use of DAT scores in combination with the CAT scores, can provide students in their Guidance classes with important self knowledge. DAT scores have traditionally been used with the Grade 9 class to assist in making Grade 10 course choices as well as to look at future career possibilities. The VR and NA subtests as well as VR&NA and Spelling and Grammar have been the most helpful subtest results. However, Absstract Reasoning, Mechanical Reasoning and Space Relations all play a useful part in measuring individual aptitudes with related interests especially in math and science related careers.

The system of grading achievement on an A to X scale, with effort graded E to U should be reassessed. It would be worthwhile knowing the proportion of letter grades which are assigned in each class. The issue of where attitude needs to be placed - either in the academic or effort scale must also be clarified.

Rank-ordering of students does not appear to add any valuable information to the overall decision making system.

The two areas which, from the results of this study, require the most attention are teacher-made test construction and teacher biases as they influence curriculum measures, marking systems and staff-student relationships. These two areas would provide a creative administrator and professional development committee with an excellent base from which to plan the year's activity.

During 1981-82, all schools in the Yellowhead School Division had one afternoon a month without students for the purpose of conducting a staff meeting. Jubilee had traditionally used this time for staff professional development.

Using Brophy & Goods' (1974) implications from the research on teacher expectations and Cooper's (1979) model for identifying expectation-effect prone teachers, an excellent individual self-awareness program could be developed. This process would blend nicely with the concepts of the advisory system which the school had begun working on during the end of the 1980-81 school year.

Further research could be carried out after the staff has received a test-construction inservice to see if correlations between the four core area subjects show independence of marks.

REFERENCES

Ahmann, J.S., & Glock, M.D. (1981). Evaluating student progress (6th ed.). Boston: Allyn & Bacon.

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Collier Macmillan.

Bachor, D.G. & Summers, G. (in press). The Canadian achievement test and the Canadian tests of basic skills: A critical, comparative review. Special Education.

Bennett, G.K., Seashore, H.G. & Wesman, A.G. (1974). Differential Aptitude Tests: Manual (5th ed.) New York: Psychological Corporation.

Brophy, J.E., & Good, T.L. (1974). Teacher-student relationships: Causes and consequences. New York: Holt, Rinehart and Winston.

Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). Evaluation to improve learning. New York: McGraw-Hill.

Cavalier, J.E. (1972). The guidance counselor's position on rank-in-class: Compared to what? A negative view of class rank. National Association of Secondary School Principals Bulletin, 56(365), 11-15.

Cooper, H.M. (1979). Pygmalion grows up: A model for teacher expectation communication and performance influence. Review of Educational Research, 49(3), 389-410.

Cronbach, L.J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper & Row.

CTC, McGraw-Hill Ryerson (1981). Canadian achievement tests: Examiner's manual. Scarborough, Ontario: Canadian Test Centre, McGraw-Hill Ryerson.

CTC, McGraw-Hill Ryerson (1983). Canadian achievement tests: Technical bulletin. Scarborough, Ontario: Canadian Test Centre, McGraw-Hill Ryerson.

Ebel, R.L. (1979). Essentials of educational measurement (3rd ed.) Englewood Cliffs, N.J.: Prentice-Hall.

Ebel, R.L. (1980). Practical problems in educational measurement. Lexington, MA: Heath.

Gronlund, N.E. (1981). Measurement and evaluation in teaching (4th ed.). New York: Macmillan.

Gronlund, N.E. (1982). Constructing achievement tests (3rd ed.). Englewood Cliffs, N.J.: Prentice Hall.

Guilford, J.P. (1954). Psychometric methods (2nd ed.). New York: McGraw Hill.

Hieronymous, A.N., Lindquist, E.F., Hoover, H.D., & King, E.M. (1978, 1979, 1982). Canadian tests of basic skills: Teacher's guide. Toronto, Ontario: Nelson, Canada.

Hills, J.R. (1981). Measurement and evaluation in the classroom (2nd ed.). Columbus, OH: Merrill.

Hotard, S.R. (1983). The stability and validity of academic achievement (Report No. CG 016 878). Paper presented at the annual meeting of the Southwestern Psychological Association: San Antonio, TX. (ERIC Document Reproduction Service No. ED 234 284).

Kelleher, P. (1982). To rank or not to rank: A solution to the class rank dilemma. The College Board Review, 124, 16-18.

Linn, R.L. (1983). Testing and instruction: Links and distinctions. Journal of Educational Measurement, 20(2), 179-189.

Macdonald, J.B. (1975). Some moral problems in classroom evaluation/testing. Urban Review, 8(1), 18-27.

Newcomer, E.K. (1972). What does RiC really mean? National Association of Secondary School Principals Bulletin, 56(365), 16-18.

Nunnally, J.C. (1975). Introduction to statistics for psychology and education. New York: McGraw-Hill.

Pedulla, J.J., Airasian, P.W., & Madaus, G.F. (1980). Do teacher ratings and standardized test results of students yield the same information? American Educational Research Journal, 17(3), 303-307.

Rosenthal, R.,& Jacobson, L. (1968). Pygmalion in the classroom: Teacher expectation and pupils' intellectual development. New York: Holt, Rinehart & Winston.

Shertzer, B.E., & Linden, J.D. (1979). Fundamentals of individual appraisal: Assessment techniques for counselors. Boston: Houghton Mifflin.

Terwilliger, J.S. (1977). Assigning grades - Philosophical issues and practical recommendations. Journal of Research and Development in Education, 10(3), 21-39.